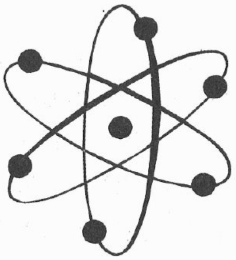
二十一世紀的新興科學

生物資訊學



◎農試所農藝系　呂秀英、呂椿棠、魏夢麗
農場室　陳烈夫

一、前　言

無庸置疑地，資訊技術和生物技術是21世紀最具發展潛力的兩大產業。資訊技術發展甚為快速，電腦和網際網路（Internet）並已滲透到社會生活之各個角落，成為當今世界資訊交流的一項重要工具和手段。生物技術則從20世紀之蓄勢待發，到今日之蓬勃發展，從基因出發，對生命科學各個層次的研究，包括基因組、蛋白質組、細胞、生物複製體等將為無數的商機打開大門。21世紀是生命科學的時代，也是訊息時代。生物資訊學（Bioinformatics）是以數學、統計理論和資訊科學的觀點、理論及方法去研究生命現象、組織和分析呈指數增長的生物資料的一門科學。生物資訊學集生命科學與資訊產業兩大本世紀最具活力的高科技領域於一體，必將大有作為，其中的商業前景極為可觀。生物資訊學與傳統育種方法相結合來提高動植物品種選育效率、創造遺傳育種資源、加快育種過程，將是育種界的必然發展趨勢。

二、解讀「生命天書」

隨著人類基因組計畫的實施，破解人類及多種模式生物的遺傳密碼已成為生物學領域的重要學科。基因組（genome，或稱基因體）是生物體所有基因的總和。所有生物的遺傳密碼都藏在DNA（去氧核糖核酸）裡。DNA已被證明是遺傳的基本物質。它以四個字母（A、C、G、T）的密碼形式，建構了生命奧秘的藍圖。在人的每個細胞中，包含人類全部遺傳訊息的24條染色體（22對體染色體及X,Y性染色體）的DNA，就是由30億個這四種不同鹽基組成；遺傳訊息的秘密，就在於這四種鹽基的有序排列。基因在細胞中以DNA雙螺旋鏈的形式存在，把DNA拉直來看，就像一串很長很長的由四個字母組成的無聊文字：-A-G-T-A-A-C-G-C-T-T-A.....，這序列在適當的時候被解碼翻譯，然後細胞按照其所提供的資訊和指令來拼裝特定的蛋白質，這些特定的蛋白質通過和生物體內其他物質的相互作用而產生各種性狀和行為。因此，基因是生命科學的核心與起點。我們可以將四個鹽基字母逢機排列組成的基因看成是生命的語言。

目前人類基因組大規模定序工作已接近尾聲，除此國際水稻基因組序列分析計畫所獲成果也急起直追，雖然不少媒體都使用了「破解生命之謎」、「解讀生命天書」等詞來讚揚科學界在這方面的成就，但事實上，完成基因組定序，僅僅是在破解生命之謎的道

作者：呂研究員秀英
聯絡電話：04-23302301轉125

路上邁出了一小步，艱辛而繁重的研究還在後頭。接下來要做的工作是所謂後基因組時代（強調的不只是生物科技，而且是跨領域的結合）的任務，即收集、整理、檢索和分析序列中表達蛋白質結構與功能的訊息，找出規律。

比方來說，構成英語的基礎是26個字母，這些字母構成單詞、句子乃至文章。從單字到文章，在不同層次上，它們都或多或少含有「訊息」。但將這26個字母胡亂排列，就不可能含有任何訊息，而只是堆「文字垃圾」。人類基因組計畫就是測定全部30億個鹽基的排列次序。以每個鹽基作為一個字母來印刷，就相當於印刷3,000本每本1,000頁每頁1,000字的「天書」！完成定序，只不過是我們打開了這本天書，但我們仍然不能讀懂它，這裡面包含的訊息太大了。我們必須在從這些定序字母中，解讀出哪些字母可以組成有意義的「實詞」，而哪些可能是「虛詞」，以甚麼樣的文法組成有意義的句子和段落，裡面包含哪些標點符號，最終組成一篇孕育和傳遞生命的宏文鉅著。所以要從基因組的全部鹽基序列獲得完整的遺傳訊息，還有漫長的路要走，而唯有掌握這些訊息，才算學會了「遺傳語文」，讀懂了這本「天書」。

三、生物資訊數據的爆炸性增長

在人類基因組解讀計畫的推動之下，決定DNA序列的實驗技術有了重大的突破，基因資料庫的資料正以指數的速率快速成長：以GCG（Genetics Computer Group）的資料庫為例，1997年8月的97-5版資料量是8.04GB，1998年3月的98-2版資料量已經膨脹到10.15GB了；而1982年12月GenBank之登錄序列有606筆，全長約68萬鹽基，到2000年6月為止，登錄序列總計7,077,491筆，全長超過86億鹽基。由於遺傳密碼的DNA是以一度空間的線性順序排列，而執行生化反應功能來維繫生命的是蛋白質，而蛋白質功能的正常與否，則是取決於三度空間的摺疊(folding)是否正確。因此除了基因序列的分析之外，另外一個重要的課題便是蛋白質的分析，包括三度空間的立體結構預測、蛋白質模組（motif）的分析與預測等。目前利用X-ray結晶繞射圖法決定蛋白質晶體結構，有同步輻射提供更強的光源；利用核磁共振決定蛋白質在液體中的結構，也有同位素標定的三維，甚至四維的光譜技術突破，於是蛋白質資料庫（例如PDB、SCOP、Swiss-Prot、PIR等）的資料量也有驚人的成長速率。

在人類的科學研究史中，這種科學資料的急速膨脹是史無前例的。面對這些巨大而複雜的資料，生物學家再也不可能用傳統的紙上作業的方法來解讀這些資料，運用電腦管理資料、控制誤差、加速分析過程勢在必行，因此結合了分子生物學、生物物理學、統計數學、資訊科學…等跨領域的專家們紛紛投入這方面的研究，整合出一個新興的研究領域－生物資訊學。

四、生物資訊學的任務及目標

簡單來說，生物資訊學是將基因組DNA序列資訊分析作為源頭，在獲得了蛋白質編碼區的資訊後並進行蛋白質空間結構模擬和預測，然後依據特定蛋白質的功能進行必要的藥物設計。因此，在後基因組研究時代，

基因組資訊學、蛋白質結構模擬以及藥物設計將是生物資訊學的三個重要組成成份。目前生物資訊學的研究範圍，大致可分爲以下任務及目標：

㈠基因組相關資訊的收集、儲存、管理及提供：建立資料庫是儲存基因組相關資訊的重要步驟，目前在網際網路上最常被檢索利用的資料庫有GenBank（美國）、EMBL（歐洲分子生物研究室聯盟）和DDBJ（日本）之全球三大DNA資料庫，以及PDB、PIR等蛋白質資料庫。將龐大的原始資料依其不同的特性予以加工，而建構出加值資料庫，不但能替使用者先處理大量資訊及進行分析計算以節省資源，同時讓使用者能集中心思解決問題而非學習使用電腦，更重要的是專業人員注入的知識會對使用者有很大的幫助。著名的二級資料庫有蛋白質結構分類資料庫(SCOP)、受體資料庫、複製載體資料庫等。而資訊整合則仰賴網際網路讓使用者和資料庫間能迅速有效地傳遞及更新資訊。透過各種資料庫的檢索和連結，對基因和蛋白質間的關係，以及各種蛋白質的機能等，可以充分且有效地被解析及探討。

㈡新基因的發現與鑑定：表現序列標幟（expressed sequence tag, EST）是從基因截錄表現的短cDNA序列，它們攜帶著完整基因某些片段的資訊。由於EST序列中包括了大量未發現的人類基因資訊，因此如何利用這些資訊發現新基因成了近幾年的重要研究課題。另外，從基因組DNA定序資料中確定編碼區的方法研究，最近也有一些新方法出現，例如考慮高維分布的統計方法、神經網絡方法、分形方法等。利用密碼學方法用於識別編碼區，也獲得不錯的結果。

㈢非編碼區資訊結構分析：非編碼區（“Junk” DNA)佔據生物基因組的大部分，雖然人們還不清楚它們的作用，但從生物進化的觀點來看，這部份序列必定具有重要的生物功能。因此尋找這些非編碼區的特徵、訊息調節與表現規律是未來相當長時間內的熱門課題。另外，由於目前對於三聯體編碼方式的區分度明顯不足，故尋找新的非三聯體的編碼方式亦爲生物資訊學家的重要任務之一。

㈣生物演化的研究：將物種分類進行親緣樹（phylogenetic trees，或稱演化樹）的建立是一門歷史久遠的學科，用來說明物種之間親疏遠近的關係。早期的分類原則是根據形態解剖學的知識建立起親緣樹，現在則拜分子生物學之賜，有大量生物序列資料，科學家利用多重序列排列的分析方法，建立起有分子生物學基礎的分子親緣樹。而由於蛋白質的結構決定了蛋白質的功能，因此對於蛋白質執行功能部位的三度空間結構進行比對，應該可以得到物種之間的親緣關係，並驗證以前所建構的親緣樹。

㈤完整基因組的比較研究：從越來越多的完整基因組所作的分析就能得到很多有意義的結論，例如人們就能估計最小獨立生物體至少需要多少基因才能存活，這些基因是如何使他們活起來的？具有大小相似的基因組（例如鼠和人都含有約30億個鹼基對，基因數也類似）之不同生物，爲何其間差異卻如此之大，其原因何在？因此表現型差異不但應從基因、DNA序列找原

因，也應考慮染色體組織上的差異。

(六)基因組資訊分析的方法研究：發展有效地能支援大尺度作圖與預測序列的軟體、資料庫、資料庫建置工具、遠程通訊工具，才能容易地處理日益增長的物理圖、遺傳圖和序列資訊。同時改進現有的理論分析方法，並建立快速且嚴格的多序列比較方法，為生物資訊學的當務之急。

(七)大規模基因功能表現圖譜的分析：目前基因組的研究已從結構基因組逐漸過渡到功能基因組。為得到基因表現的功能譜，國際上在核酸及蛋白質兩層次上都發展了新技術，如DNA晶片（核酸層次）、二維凝膠電泳和測序質譜（蛋白質層次）。從數學角度來看，此並非簡單的動態系統或不確定性問題，因此需要發展新的方法和工具。所以無論是生物晶片或蛋白質組技術之發展都更強烈地依賴於生物資訊學的理論、技術與資料庫。

(八)蛋白質分子空間結構的預測、模擬及分子設計：欲了解基因的功能，要找到這些蛋白質的分子基礎，必須進一步知道它們的三維結構。同時，設計藥物也要了解相應的蛋白質受體的三維結構。近年來對蛋白質結構模式的研究有很大的進展，一般公認蛋白質的摺疊類型是有限的，目前估計為幾百至幾千種，但這還小於蛋白質所具有的自由度數目。同時蛋白質的的摺疊類型與其組成分和一級序列相關，這樣就有可能從蛋白質的初級訊息中確定它們的最終摺疊類型。研究一個基因組中所編碼的所有蛋白質的結構、功能及相互關係，稱為蛋白質組學，這不僅是分子生物實驗問題，也是一個生物資訊學問題。

(九)藥物設計：傳統的藥物研製主要是從大量的天然產物，如動物、植物、微生物和合成有機、無機化合物中進行篩選。往往得到一個可供臨床使用的新藥物，其開發過程要耗費大量的金錢和時間。近年來由於結構生物學的發展，相當數量的蛋白質及一些核酸、多醣的三維結構已被精確了解，因此基於生物巨分子結構的藥物設計成為當前的熱門課題。生物資訊學的研究不但可提供生物巨分子空間結構的資訊，還能提供電子結構以及動力學行為的資訊，此為天然生物巨分子的改變特性和基於受體結構的藥物分子設計提供了依據，同時由模擬結果也能瞭解生命現象的基本過程。

(十)應用與發展研究：基因組資訊學的研究成果不但具有重要的理論價值，也可直接應用到農業生產和醫療中。隨著人類基因組、水稻基因組及各種模式生物基因組的解釋，根據不同物種間的進化距離和功能基因的同源性，可以容易地找到各種家畜、經濟作物與其經濟效益相關的基因，並進而對它們按照人們的願望加以改造。而分子生物學常用的表現載體、聚合酶連鎖反應(PCR)引子及各種試劑模組的設計也必須依賴於核酸的序列資訊，故基因組資訊學提供的大量資訊為這類技術之發展提供了廣闊的天地。

五、生物資訊學孕育著巨大的商機

生物資訊學的發展將會對生命科學帶來革命性的變革。藉助於基因資料庫的內容，經過生物資訊學對DNA和蛋白質的序列及結構進行收集、整理、儲存、發表、提取、加

工、分析及研究，從大量不連貫的資訊中發現隱藏的新線索、新現象和新規律，從而找到各種有用的新基因、新蛋白和其功能。科學家便可以利用生物密碼製造出任何他們可以想像得到的生物巨分子，用來做爲藥物，可以改良動植物，增加糧食產量等，爲人類造福。它的成果不僅對生物醫學、農學、遺傳學、細胞生物學等相關基礎學科起強大的推動作用，而且還將對醫藥、分生、食品、農業等產業及環境監測產生巨大的影響，甚至引發新的產業革命。

目前各國政府、科研單位、工業界、全球各大藥廠對此皆極爲重視，莫不競相投入大量資金與人力，發展分析軟體，甚至建立自己獨有的資料庫。加值資料庫是未來生物學研究必備的工具。加工後的資訊，往往比產生這些資訊的原有的生意更有價值。迄今生物科技公司和製藥工業等以資訊爲基礎而形成公司內部的生物資訊學部門的數量與日俱增。每天都會有新的生物資訊公司及研究機構生物資訊中心產生。而歐美各國、日本及中國也都相繼成立了國家級生物資訊資料中心。當前這些生物資訊中心的發展重點及具體成果主要在於分析軟體的形成、資料庫的建立、網際網路資訊搜尋和導航系統等各方面。它爲生物產業提供資料和訊息服務。通過網際網路，爲客戶提供有償服務是生物資訊學商業化模式的一種。例如美國最有名的Celera公司在網路上進行基因資料的收費查詢和分析就是一個例子。生物資訊學產業的價值在1998年已經達到18億美元，而到2002年估計可增長至2000億美元以上，這是目前任何國家政府的科技決策者都不能視而不見的。美國將考慮批准投資160億美元，在美國建立5至20個生物資訊學中心。

美國衛生部研究計畫審查委員兼范德堡大學醫學院癌症中心生物統計主任石瑜指出，生物資訊自1991年出現至今，不到十年時間，但卻因人類數以萬計的基因組，需要大量資訊統計與機率分析，歐美各國都已起步發展中，他認爲，臺灣電腦軟體與系統已建立基礎，生物資訊統計產業是不錯的發展方向。

六、培育生物資訊人才為當務之急

生物資訊學的特點是投資少、見效快、效益大，適合於我國的現實條件。關鍵在於有關學科之間的密切合作和加速培養生物資訊專業人才。這樣的人才當前全世界都十分缺乏。生物資訊學是跨多學科的學問，涉及生物、數學、統計、物理、化學、資訊科學等傳統領域，目前尚處於初期發展的階段，還是一個相對較新的領域，因而大部分教育機構仍未將其列入常規課程中。而伴隨著生物資訊機構的快速增長，生物資訊學人才將成爲搶手貨。國際上生物資訊人才的需求愈來愈大，供不應求的情況也會持續。再不改進，缺乏生物資訊學人才將是發展基因科技的主要瓶頸之一。符合資格的生物資訊人才需要經過長時間的適當培訓才可達到，但爲緩不濟急，絕大多數生物資訊中心正積極從資訊管理、生物統計及生命科學等相關學科之學術界徵募人才，以尋求有能力的替代人手。這是因爲生物資訊的取得、收集、整理、除錯、建檔和分析都需要相當能力的生物統計及資訊處理人才。

目前我國已認識到生物資訊學的重要性，各研究單位及政府機關也開始積極投入這個新興領域，陸續組成跨領域的研究群，像是國家衛生研究院、中央研究院、台北榮

民總醫院教研部、國家高速電腦中心、陽明大學、清華大學、臺灣大學生物技術研究中心等都設有生物資訊網站，並紛紛開設了生物資訊學課程及培訓班。國科會也自2000年起對外公開徵求生物資訊學研究計畫，希望能經由該會的推動，使得我國生物科技的發展能在國際上佔有先機，並儘快培養國內生物資訊的研究及相關人才。但與國外相較，我國科研單位投入生物資訊學領域的研究人員仍嚴重不足，這主要是缺乏系統教育，故成效甚微。國內的生物統計學專家為數不多且零星散佈於各大學與研究單位之中，大部分所從事的都是生物統計理論研究與教學，真正應用在生物統計諮詢及實務操作的研究者屈指可數，因此如何充分發揮現有少數人才和單位的潛力，優勢互補，讓更多研究工作者加入生物資訊學的領域共同打拼，是刻不容緩之事。而具體的辦法應該朝向「生物統計與資料處理中心」或「生物資訊中心」的整體架構進行，設置一常態性的「研究服務」支援單位，如此才能在國際上搶得目前最熱門的市場。對從事研究服務的研究人員，如生物統計學家更應賦予必要的支援。

對生物統計學家而言，迎接21世紀新科技蓬勃發展的時代，自有其不可抗拒的使命，但也應對自己所扮演的角色做一番自我調適，除了本身的數理背景及獨立執行資料分析能力以外，更需要有基本的生物醫農背景，同時必須擁有電腦資訊、資料庫及資料處理的理論及實際操作能力。而生物醫農專家在展望生物統計專家在對生物資訊的可能具體貢獻之前，也應有一前瞻性的瞭解：統計分析絕非一般技術性的例行性工作，由於生物科技是一項必須投入相當人力、物力與財力的浩大工程，試驗計畫事前的謹慎評估、試驗設計和資料收集的方法與步驟研擬、試驗結束後的資料處理與分析，以及最後結果的解釋等，都是研究者必須借重生物統計專家的地方。否則浪費了許多無謂的資源去做了一些可能沒有結果的研究，是非常可惜的。單打獨鬥、關起門來做研究的時代應該已經過去，唯有接受「研究合作」的觀念，才能使生物醫農研究的推展更順利。

七、參考文獻

李曉、雷波。1999。生物信息學理論及農業應用的發展趨勢。西南農業學報 12（增刊2）：32-35。（中國）

陳潤生。1999。當前生物信息學的重要研究任務。生物工程進展 19(4)：11-14。（中國）

輕部征夫。2000。生物情報學。技術 と經濟 399：16-29。（日本）

趙雅婷。2000。生物資訊學之發展與應用。近代作物科學發展研討會論文集，pp20-27，臺灣大學農藝系，台北，中華民國。

Attwood, T. K. and D. J. Parry-Smith. 1999. Introduction to Bioinformatics. Addison Wesley Longman Limited, UK, 218P.

Gershon D., B. W. Sobral, B. Horton, P. Wickware, H. Gavaghan and M. Strobl. 1997. Bioinformatics in a post-genomics age. Nature 389: 417-422.

Ouzounis, C. 2000. Two or three myths about bioinformatics. Bioinformatics 16: 187-189.

Ruediger, N. 1996. Bioinformatics: New frontier calls young scientists. Science 273:265.